STATISTICS IN MEDICINE, VOL. 11, 1057–1063 (1992)

# COMPARISONS BETWEEN A NEW DRUG AND ACTIVE AND PLACEBO CONTROLS IN AN EFFICACY CLINICAL TRIAL

CHARLES W. DUNNETT

*Department of Mathematics & Statistics, and Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, Ontario L8S 4K1, Canada*

AND

AJIT C. TAMHANE

*Department of Statistics, and Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, Illinois 60208, U.S.A.*

## SUMMARY

D'Agostino and Heeren (DH) discussed the multiple comparison issues that arise in evaluating both sensitivity and efficacy in over-the-counter drug trials. We propose a general definition of sensitivity that includes DH's definition as a special case. We also propose a test for sensitivity that coincides with the MIN test of Laska and Meisner at one extreme but has the advantage of identifying specific drugs satisfying the sensitivity criterion when some fail to do so. We suggest that the test of Dunnett as well as an extension of it may be useful for the efficacy comparisons.

## 1. INTRODUCTION

D'Agostino and Heeren[1] (referred to as DH in the sequel) have discussed some issues arising in a clinical trial involving multiple comparisons between a new drug, one or more established drugs as positive controls and a placebo as a negative control. The positive controls serve as reference standards for efficacy comparisons with the new drug while the placebo is used to validate the clinical trial by confirming that the known active drugs are effective relative to the placebo.

In their paper, DH argued in favour of employing statistical tests that control experimentwise rather than comparisonwise error rates in making the multiple comparisons between treatments. Although we agree with the experimentwise approach, we disagree with some aspects of DH's reasoning which led them to advocate the use of 'Dunnett's procedure'[2] to establish the 'downside' sensitivity of the clinical trial. We will show that, when sensitivity is defined to require that *all* actives are more effective than placebo, the correct procedure for controlling experimentwise error uses (in effect) comparisonwise tests. (This point was also made by some of the discussants of the DH paper.) The MIN test of Laska and Meisner[3] is thus the correct test.

We propose, however, a more flexible test which coincides with MIN if all the active drugs meet the sensitivity criterion but has the advantage that it identifies the particular active drugs that the investigator can demonstrate to be effective when some fail to meet the criterion. The proposed test is the step-up (SU) test procedure developed by Dunnett and Tamhane.[4]

0277–6715/92/081057–07$05.00


*Received April 1991*
*Revised December 1991*

On the other hand, if the objective is to show that *at least one* of the active drugs is effective, then the procedure described in Dunnett[2] is appropriate as a single-step (SS) test. As we will discuss, this objective is more realistic in the efficacy comparisons between the test drug and active drugs than in the placebo comparisons to test sensitivity. A step-down version of SS, denoted by SD (see Dunnett and Tamhane[5]), achieves higher power than SS for detecting individual effects while maintaining the desired experimentwise error rate control. A possible disadvantage of these stepwise tests is that they do not have simultaneous confidence interval counterparts.

To clarify some of the issues involved, we propose in the next section a general definition of the concept of sensitivity of a clinical trial that includes DH's definition as a special case. We then discuss three problems faced by the investigator, in the following order: (1) testing for sensitivity, (2) establishing the efficacy of the test drug, and (3) comparing the efficacy of the test drug to the positive controls. As DH emphasized, it is necessary to establish (1) and (2) before considering (3). In the final section, we provide a justification for treating these three problems separately rather than simultaneously as some discussants of DH proposed. We also discuss other matters, such as the need to check assumptions and the choice of statistical tests when some of the usual assumptions are not met.

## 2. TESTING FOR 'DOWNSIDE' SENSITIVITY

DH state that the investigator must demonstrate all the active drugs to be more effective than the placebo to have a valid trial for testing efficacy. In their rejoinder to the discussants of the paper, however, DH back off from this stand to some extent and appear to concede that in some circumstances it may not be necessary to show that all of the actives are effective. Hence, we propose a general definition of sensitivity that permits some flexibility, as follows. Denote by $k$ the number of active drugs in the experiment ($k \geqslant 1$). We define sensitivity to mean that the investigator shows at least $m$ of the $k$ active drugs are effective compared with the placebo, where $m$ is a number $\leqslant k$ specified in advance.

Denote the test drug by A, the positive controls by $B_1, \ldots, B_k$ where $k$ can be any number (for instance, $k = 2$ in the example discussed by DH) and the placebo by P. Denote the expected responses for A, $B_i$ and P by $\mu_A$, $\mu_i$ and $\mu_0$, respectively. Then the parameters of interest for the purpose of establishing sensitivity are $\mu_i - \mu_0$ ($1 \leqslant i \leqslant k$). Let $H_{0i}$ denote the hypothesis that the $i$th drug does not meet the sensitivity criterion and $H_{1i}$ the hypothesis that it does. Also, denote by $t_i$ the test statistic used to test $H_{0i}$ versus $H_{1i}$. (For example, when the assumptions of homogeneous variances and normality are met, $t_i$ is the usual Student $t$ statistic with an overall pooled error estimate.)

Two special cases of the general definition of sensitivity are of particular interest. In the following, we assume that larger values of the efficacy variable are associated with greater efficacy; in a trial where smaller values are better, the direction of certain inequalities must be reversed.

**Case 1:** $m = k$

This corresponds to the definition proposed by DH, requiring that the investigator show all of the $k$ active drugs are effective to have a valid trial. Since the sensitivity criterion fails if just one of the active drugs is ineffective, the null hypothesis tested and its alternative are

$$H_0\colon\ \mu_i - \mu_0 \leqslant 0 \text{ for at least one } i$$

versus

$$H_1\colon\ \mu_i - \mu_0 > 0 \text{ for all } i = 1, \ldots, k.$$

Rejection of $H_0$ establishes sensitivity. The hypothesis $H_1$ is the intersection of the individual $H_{1i}$ and the hypothesis $H_0$ is the union of the $H_{0i}$, hence this is called an *intersection-union* problem; see Berger.[6] The test consists of rejecting $H_0$ if *all* the $t_i$'s exceed a certain critical value, $c$. Equivalently, the test consists in ordering the statistics $t_i$ to form a non-decreasing sequence $t(1) \leqslant \ldots \leqslant t(k)$ and rejecting $H_0$ if $t(1) = \min t_i$ exceeds $c$.

The critical constant $c$ should be chosen so that the probability of a type I error is $\leqslant \alpha$ for all parameter values consistent with $H_0$. Berger[6] has shown that the maximum value achieved by this probability occurs when $\mu_i - \mu_0$ equals zero for exactly one $i$ and approaches $+\infty$ for all the others. Hence the correct critical value for testing $t(1)$ is the usual $\alpha$-point of Student's $t$. This is exactly the Laska and Meisner[3] MIN test. Thus the intersection-union problem is solved by using level-$\alpha$ comparisonwise tests for the component hypotheses.

However, suppose the test does not reject $H_0$. This means that at least one of the active compounds has failed the sensitivity criterion. We feel it is wrong simply to stop there and declare the trial invalid, since there may be an alternative explanation. For example, some form of data loss such as dropouts may have decreased the sample size in a particular group resulting in a decrease in the power for that particular comparison, but without necessarily invalidating the other comparisons. If so, it might still be useful to test whether some of the active drugs are effective.

The MIN test is not appropriate for this purpose, as it does not apply to the alternative at issue and its critical value would not satisfy the experimentwise error rate criterion. However, an appropriate method is Dunnett and Tamhane's[4] SU test. This method was developed for testing a set of equicorrelated contrasts and applies to the present problem when the data are balanced (that is, equal sample sizes for the $k$ active drugs, with possibly a different sample size for the placebo). The method uses an increasing sequence of critical values $c_1 < c_2 < \ldots < c_k$ to apply to $t(1), t(2), \ldots, t(k)$ in step-up fashion, starting with $t(1)$ and continuing as long as $t(i) < c_i$. When $t(i) \geqslant c_i$ for the first time, one can declare all the actives, corresponding to that $i$ and higher values, superior to the placebo. Thus, we can establish sensitivity in terms of these actives and this might salvage some useful information from the trial.

The critical constants $c_i$ must be determined such that the SU test satisfies the experimentwise error rate criterion. (See Table I for a short table of both one-sided and two-sided $c_i$ values for $k \leqslant 5$. They apply when all sample sizes are equal, including the control, which makes the correlation coefficients between the contrasts equal to the common value $\frac{1}{2}$.) The first critical

Table I. Critical values for step-up (SU) procedure for balanced data case (equal sample sizes)

| | | 1-sided | | | | | 2-sided | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | d.f. | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ |
| 0·05 | 10 | 1·81 | 2·17 | 2·35 | 2·47 | 2·57 | 2·23 | 2·59 | 2·77 | 2·90 | 2·99 |
| | 20 | 1·72 | 2·05 | 2·20 | 2·31 | 2·39 | 2·09 | 2·39 | 2·55 | 2·66 | 2·74 |
| | 30 | 1·70 | 2·01 | 2·16 | 2·26 | 2·34 | 2·04 | 2·33 | 2·48 | 2·58 | 2·66 |
| | 60 | 1·67 | 1·97 | 2·11 | 2·21 | 2·29 | 2·00 | 2·28 | 2·42 | 2·51 | 2·58 |
| | $\infty$ | 1·65 | 1·93 | 2·07 | 2·17 | 2·24 | 1·96 | 2·22 | 2·35 | 2·44 | 2·51 |
| 0·01 | 10 | 2·76 | 3·13 | 3·32 | 3·46 | 3·56 | 3·17 | 3·54 | 3·75 | 3·89 | 4·00 |
| | 20 | 2·53 | 2·82 | 2·98 | 3·09 | 3·17 | 2·85 | 3·13 | 3·29 | 3·40 | 3·48 |
| | 30 | 2·46 | 2·73 | 2·88 | 2·98 | 3·05 | 2·75 | 3·02 | 3·16 | 3·26 | 3·33 |
| | 60 | 2·39 | 2·65 | 2·78 | 2·87 | 2·94 | 2·66 | 2·90 | 3·03 | 3·12 | 3·19 |
| | $\infty$ | 2·33 | 2·56 | 2·69 | 2·77 | 2·84 | 2·58 | 2·80 | 2·92 | 3·00 | 3·06 |

constant $c_1$ is the ordinary $\alpha$-point of Student's $t$ and is identical with the critical constant used for MIN, so the two tests have identical requirements for establishing whether the trial is valid. The only difference between them is that SU identifies which actives do and which do not satisfy the sensitivity criterion, when it is not possible to demonstrate strict sensitivity according to the definition.

**Case 2:** $m = 1$.

This corresponds to defining the sensitivity criterion to require that at least one active drug is effective. The null hypothesis tested and its alternative are:

$$H_0\colon\ \mu_i - \mu_0 \leqslant 0 \text{ for all } i,$$

versus

$$H_1\colon\ \mu_i - \mu_0 > 0 \text{ for at least one } i.$$

This is Roy's *union-intersection* problem (see Hochberg and Tamhane,[7] page 28), since $H_1$ is the union of the separate $H_{1i}$ and $H_0$ is the intersection of the $H_{0i}$. The test statistic for $H_0$ is $t(k)$, the largest of the $t_i$'s instead of $t(1)$ as in the previous case. The test rejects $H_0$ if $t(k)$ exceeds a critical value, $c'$.

In this case, the type I error probability is maximized when $\mu_i - \mu_0 = 0$ for all $i$, which means that we should choose $c'$ as the upper $\alpha$-point of $k$-variate $t$ to achieve a type I error rate $\leqslant \alpha$. Moreover, we can reject the individual $H_{0i}$ if the corresponding $t_i \geqslant c'$ and this multiple comparisons test controls the experimentwise error rate at level $\alpha$ (Hochberg and Tamhane,[7] Chapter 2, Section 2.1). This is, in fact, the basis for the test in Dunnett[2] and is the single-step test denoted by SS in Dunnett and Tamhane.[4]

A test that has greater power than SS, however, is to proceed stepwise using a sequence of critical values $c'_k, c'_{k-1}, \ldots$ to test $t(k), t(k-1), \ldots$, starting with $t(k)$ and continuing as long as $t(i) \geqslant c'_i$, indicating a rejection of the corresponding separate hypothesis, and stopping when $t(i) < c'_i$ indicating that and any remaining hypotheses cannot be rejected. This test is step-down (denoted by SD for brevity) and has a long history; see Dunnett and Tamhane.[5] We can think of SU and SD as stepwise versions of the MIN and SS tests, respectively.

**General case:** $1 \leqslant m \leqslant k$.

For values of $m$: $1 < m < k$, we express the null hypothesis $H_0$, stating that the sensitivity criterion is not met, along with its alternative $H_1$ as:

$$H_0\colon\ \mu_i - \mu_0 \leqslant 0 \text{ for at least } k - m + 1 \text{ values of } i,$$

versus

$$H_1\colon\ \mu_i - \mu_0 > 0 \text{ for at least } m \text{ values of } i.$$

The appropriate statistic for a single-step test is the $m$th largest of the $t_i$'s or $t(k-m+1)$. It can be shown that the critical value for $t(k-m+1)$ is the $\alpha$-point of multivariate $t$ in $k-m+1$ dimensions. Note that this yields the SS test for $m = 1$ and the MIN test for $m = k$.

Alternatively, we can use the step-up test SU as before, starting with $t(1)$ and stopping when $t(i) \geqslant c_i$. If we stop by the $(k-m)$th step, the number of actives found effective is sufficient to satisfy the sensitivity criterion. We could also use the step-down SD test, starting with $t(k)$ and terminating when $t(i) < c'_i$; we meet the sensitivity criterion if it takes $m$ steps or more.

The special cases 1 and 2 should cover most practical needs. In particular, case 1 is likely to be more appropriate for testing downside sensitivity. The MIN test can be used as a single-step test, but the stepwise SU test is preferable since it identifies which active drugs are effective.

## 3. ESTABLISHING THE EFFICACY OF THE TEST DRUG

To establish the efficacy of the test drug, it is necessary to show that $\mu_A - \mu_0 > 0$. This is identical in form to the requirements on the individual active drugs considered in establishing the sensitivity of the trial, but with the distinction that a failure to establish the efficacy of the test drug relative to the placebo reflects on the test drug, rather than on the validity of the experiment. For this reason we consider it important to treat this difference separately from the differences between the known actives and the placebo. After establishing the validity of the experiment with respect to sensitivity and before proceeding with the efficacy comparisons, the next question is whether we can show the new drug to be effective relative to the placebo.

The null hypothesis to be tested and its alternative are

$$H_{0A}: \mu_A - \mu_0 \leqslant 0 \quad \text{versus} \quad H_{1A}: \mu_A - \mu_0 > 0.$$

Since this is a single null hypothesis, we can use the usual $t$-statistic and apply a comparisonwise level-$\alpha$ test.

## 4. COMPARING THE EFFICACY OF THE TEST DRUG WITH THE POSITIVE CONTROLS

The separate null hypothesis and its alternative for testing whether A is more efficacious than $B_i$ is expressed as

$$H_{0i}: \mu_A - \mu_i \leqslant 0 \text{ versus } H_{1i}: \mu_A - \mu_i > 0.$$

Note that, although we have written this as a one-sided hypothesis testing problem, we could also formulate it as two-sided. Certainly, the one-sided alternative as we have written it expresses the direction of the difference that the sponsor of the test drug hopes to establish. On the other hand, a difference in the other direction might also be of interest to a regulatory agency who may require assurance that the new drug is not worse than any of the standard drugs. Furthermore, there is no *a priori* reason to expect the differences to occur in a specific direction, as there was with the placebo comparisons. Thus, a two-sided test might be more appropriate, which is approximately equivalent to two one-sided tests each at level $\alpha/2$. (The reader is referred to Peace[8] and other papers in the same journal for a discussion of the various points of view on one-sided and two-sided testing.) The principles are unchanged, however, whether we use one-sided or two-sided tests; for the purposes of the present discussion we leave it in its one-sided form.

These hypotheses for $i = 1, \ldots, k$ are of the same form as the separate hypotheses considered in the downside sensitivity problem, the only change being that the common element in each $H_{0i}$ is now $\mu_A$ whereas previously it was $\mu_0$. The first question is whether we should consider the $H_{0i}$'s separately or simultaneously. One could make an argument in favour of a separate approach if they involved $k$ separate questions (see O'Brien[9]). This would be valid if, for example, the sponsor planned to run separate advertising campaigns for the new drug comparing it with each of the competing $B_i$'s, but this hardly seems likely. It seems more likely that the investigator plans to use the current trial to identify the particular active drugs in comparison with which the test drug appears to have the greatest efficacy (and choose the marketing strategy for the new drug accordingly). If so, a simultaneous test is indicated, and then the same considerations as were discussed in the downside sensitivity problem arise. In particular, we have to decide whether the simultaneous inference problem concerns the establishment that the test drug is superior to *all* of the positive controls or, at the other extreme, superior merely to *at least one* of them. In the case of the downside sensitivity problem, we could make a good argument for requiring the establishment of *all* comparisons as non-null, since the positive controls are known actives that must

differ from the placebo. We are not likely to have similar prior evidence available, however, that would justify the superiority of the new drug over all of its potential competitors. Nor does it seem reasonable for a regulatory authority to require the superiority of a new drug to all currently available drug treatments. Evidence that it is superior to at least one of the standard drugs should suffice to establish that it merits having a place on the market and at the same time provide the sponsor with a claim which they can advertise. (This assumes, of course, that all other issues such as toxicity have been resolved satisfactorily.)

Hence, for the efficacy comparisons, a reasonable goal is to determine whether one can show that the new drug has greater efficacy than one or more of the positive control drugs. This falls under Case 2 above. Thus, the appropriate test is either the SS (single-step, see Dunnett[2]) or, for higher power, the SD (step-down, see Dunnett and Tamhane[5]) extension of the SS test.

Another possibility, proposed by DH and discussed by Koch,[10] is to formulate the goal as establishing that the test drug is equivalent or better than *all* of the positive control drugs. Here the separate hypothesis and its alternative for determining if the test drug is at least equivalent to the $i$th standard drug are

$$H_{0i}: \mu_A - \mu_i \leqslant -d \quad \text{versus} \quad H_{1i}: \mu_A - \mu_i > -d,$$

where $d$ is a specified threshold difference defining clinical importance. This simultaneous inference problem falls under Case 1 above, so the appropriate tests are ordinary $t$-tests (or MIN), as discussed in Section 2. This approach might appeal to a regulatory agency that wished to ensure that any new drug actually inferior to one of the available drugs would not receive approval. From the point of view of the sponsor, however, a trial that merely showed the new drug as no worse than its competitors would not provide much promotional material for its advertisement as a new product. An approach that might satisfy both would be first to test for equivalency of the test drug relative to the standards and then to test that its efficacy is actually superior to one or more of them. Another aim might be to show the test drug is equivalent to the standards in terms of efficacy and better than one or more of them by having less toxicity.

## 5. DISCUSSION

First, we consider the justification for the separation of the treatment comparisons of interest in the clinical trial into the three families concerned with (1) sensitivity, (2) efficacy relative to placebo and (3) efficacy relative to the positive controls, and the use of a separate experimentwise (or familywise, FWE) error rate $\alpha$ for each. Since it is necessary for the investigator to establish all three of the above, namely that the sensitivity criterion is met ($m$ out of $k$, for whatever value of $m$ is appropriate) *and* that the test drug is effective relative to the placebo *and* that the test drug is more effective (hopefully) than at least one of the known active drugs, this follows immediately when we realize that they represent another example of an intersection-union problem. Thus, we achieve an overall experimentwise type I error rate $\leqslant \alpha$, according to the intersection-union principle, by testing each of the three families of tests separately using a level $\leqslant \alpha$; see also Koch[10] for a similar argument.

We have purposely omitted any mention of comparisons among the positive control drugs. Even though an investigator might have an interest in learning that competitor B's drug is more effective than competitor C's, this does not concern the main purpose of the trial, which is the establishment of the efficacy of the new drug. Hence, we must consider it a separate question (and test it with a separate family of tests). Including such comparisons in the family of comparisons concerned with the efficacy of the test drug would reduce the power of the statistical tests.

Although we have attempted to point out the potential usefulness of stepwise testing procedures for the types of multiple comparisons described by DH that arise in these trials, and to

recommend in particular the step-up (SU) and step-down (SD) methods, we stress that these are basically hypothesis testing methods. For simultaneous confidence interval estimation, one should use the SS method.

In the DH paper, the authors recommended that the first step in any statistical analysis is an overall test of treatment differences using an ANOVA. We approve of the ANOVA, but mainly because it provides a pooled variance estimate. It is also important at this stage, however, to check the assumptions of the statistical tests, such as homogeneity of the variances and normality. The rule-of-thumb that having equal sample sizes in the treatment groups makes the homogeneous variance assumption unnecessary does not apply when one pools more than two variances. (For example, consider the effect of one of the treatment groups having a larger variance than the others: use of a common pooled variance estimate would bias upwards all the $t_i$ statistics that involved that group and bias downwards all the others.) When there is doubt regarding the variance homogeneity assumption, one should consider separate variance estimates. Similarly, one should have data that are reasonably normally distributed, although this may not be an issue since for large samples normality applies approximately to the means by the Central Limit Theorem. If in doubt, one should use robust methods, such as rank tests. There are analogues of the SD and SU procedures available for use with such tests, although the step-up versions may require some development for application in specific cases.

Presently, one can use the SU method as developed by Dunnett and Tamhane[4], and be guaranteed that the FWE requirement is satisfied, only with balanced data, under normal theory and homogeneous variances. In an application that fails to meet these restrictions, one can apply another step-up procedure due to Hochberg[11] that uses Bonferroni critical values. Although slightly less powerful, it applies under more general conditions.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Charles H. Goldsmith as well as the referee and editor for suggestions that improved the reading of the paper. The first author's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

## REFERENCES

1. D'Agostino, R. B. and Heeren, T. C. 'Multiple comparisons in over-the-counter drug clinical trials with both positive and placebo controls', (with Comments and Rejoinder), *Statistics in Medicine*, **10**, 1–31 (1991).
2. Dunnett, C. W. 'New tables for multiple comparisons with a control', *Biometrics*, **20**, 482–491 (1964).
3. Laska, E. M. and Meisner, M. J. 'Testing whether an identified treatment is best', *Biometrics*, **45**, 1139–1151 (1989). (See also their Comment on DH above, 17–20.)
4. Dunnett, C. W. and Tamhane, A. C. 'A step-up multiple test procedure', *Journal of the American Statistical Association*, **87**, 162–170 (1992).
5. Dunnett, C. W. and Tamhane, A. C. 'Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts', *Statistics in Medicine*, **10**, 939–947 (1991).
6. Berger, R. L. 'Multiparameter hypothesis testing and acceptance sampling', *Technometrics*, **24**, 295–300 (1982).
7. Hochberg, Y. and Tamhane, A. C. *Multiple Comparison Procedures*, Wiley, New York, 1987.
8. Peace, K. E. 'One-sided or two-sided $p$ values: which most appropriately address the question of drug efficacy', *Journal of Biopharmaceutical Statistics*, **1**, 133–138 (1991). (In addition, see papers on this topic by other authors which are published in the same issue.)
9. O'Brien, P. C. 'The appropriateness of analysis of variance and multiple comparison procedures', *Biometrics*, **39**, 787–788 (1983). (See also Comment on DH[1] above, 9–11.)
10. Koch, G. G. 'Comment on DH[1]', *Statistics in Medicine*, **10**, 13–16 (1991).
11. Hochberg, Y. 'A sharper Bonferroni procedure for multiple tests of significance', *Biometrika*, **75**, 800–802 (1988).